Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images

Rakshith Subramanyam[*1] **Vivek Narayanaswamy**[*1] **Mark Naufel**[1] **Andreas Spanias**[1]
Jayaraman J. Thiagarajan[2]

Abstract

Inverting an image onto the latent space of pre-trained generators, e.g., StyleGAN-v2, has emerged as a popular strategy to leverage strong image priors for ill-posed restoration. Several studies have showed that this approach is effective at inverting images similar to the data used for training. However, with out-of-distribution (OOD) data that the generator has not been exposed to, existing inversion techniques produce sub-optimal results. In this paper, we propose SPHInX (StyleGAN with Projection Heads for Inverting X), an approach for accurately embedding OOD images onto the StyleGAN latent space. SPHInX optimizes a style projection head using a novel training strategy that imposes a vicinal regularization in the StyleGAN latent space. To further enhance OOD inversion, SPHInX can additionally optimize a content projection head and noise variables in every layer. Our empirical studies on a suite of OOD data show that, in addition to producing higher quality reconstructions over the state-of-the-art inversion techniques, SPHInX is effective for ill-posed restoration tasks while offering semantic editing capabilities.

1. Introduction

In the past few years, generative adversarial networks (GANs) (Goodfellow et al., 2014) have been shown to produce high-quality, photo-realistic images in a variety of image synthesis and manipulation tasks (Karras et al., 2019; Härkönen et al., 2020; Brock et al., 2019; Song et al., 2021). In particular, the StyleGAN-v2 architecture and its variants (Karras et al., 2019; 2020; 2021) have been used to synthesize very high resolution images. At a basic level, StyleGAN-v2 learns to transform a latent vector $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^{512}$ to an intermediate latent code $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^{512}$ through a mapping function (projection head) f, which is then used to synthesize images.

The continued progress in training GANs has led to a surge in techniques that can leverage deep generators as priors for ill-posed image inversion problems (Bora et al., 2017; Anirudh et al., 2019; Daras et al., 2021). In this context, the problem of accurately embedding a given image onto the latent space, often referred to as GAN inversion, has gained significant research interest (Abdal et al., 2019; 2020; Daras et al., 2021; Zhu et al., 2020b; Wulff & Torralba, 2020; Kang et al., 2021). Furthermore, it has been demonstrated that the rich semantic information encoded in the latent space of a pre-trained StyleGAN allows seamless editing of images through controlled latent code manipulations (Zhu et al., 2020b; Karras et al., 2019).

Broadly, existing approaches for StyleGAN-based inversion perform a careful selection of the latent space for optimization ($\mathcal{Z}, \mathcal{Z}+, \mathcal{W}$ and $\mathcal{W}+$) and regularization techniques. Existing inversion strategies have been successful with images that are similar to the data used for training the generator (e.g., FFHQ faces). However, embedding out-of-distribution images (e.g., 'in-the wild'/domain shifted face images) onto such latent spaces is known to be very challenging. As a result, there has been a significant emphasis on improving StyleGAN based priors for out-of distribution inversion. For instance, Kang et al. proposed an out-of-domain face image inversion strategy by introducing an encoder-based regularization on the StyleGAN-v2 feature maps. On the other hand, Abdal et al. demonstrated that it is possible to invert even non-face images, e.g., car images, onto the $\mathcal{W}+$ space of a StyleGAN pre-trained on face images. However, the perceptual quality of the reconstructed images in both cases are significantly poorer due to the choice of latent space for optimization as well as the mismatch between the latent space prior and the OOD data.

[*]Equal contribution [1]Arizona State University [2]Lawrence Livermore National Laboratories, Livermore, CA, USA. Correspondence to: Rakshith Subramanyam <rsubra17@asu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC.and was supported by the LLNL-LDRD Program under Project No. 21-ERD-012.

Table 1: StyleGAN-based inversion involves optimizing different combinations of latent spaces ($\mathcal{Z}+, \mathcal{W}+, \mathcal{S}$ and $\mathcal{B}$). A wide variety of optimization strategies have been proposed in the literature to improve the efficacy of this inversion process.

Method	Optimization Space	Additional Regularization Strategy	OOD
PGD (Karras et al., 2019)	$\mathcal{Z}$	-	-
PULSE (Menon et al., 2020)	$\mathcal{Z}+$	Latent space search with Gaussian prior	-
ILO (Daras et al., 2021)	$(\mathcal{Z}+, \mathcal{S}, \mathcal{B})$	ℓ_1-ball constraint on manifold induced by the previous layer	-
I2S (Abdal et al., 2019)	$\mathcal{W}+$	-	✓
Zhu *et al.* (Zhu et al., 2020b)	$\mathcal{W}+$	PCA whitening in $\mathcal{W}+$ space (P-norm$^+$)	-
IDInvert (Zhu et al., 2020a)	$\mathcal{W}+$	In-domain regularization using domain-guided encoder	-
PIE(Tewari et al., 2020a)	$\mathcal{W}+$	Hierarchical non-linear optimization	-
Wulff *et al.* (Wulff & Torralba, 2020)	$\mathcal{W}+$	Statistical priors on $\mathcal{W}+$ space	✓
StyleFlow (Abdal et al., 2021)	$\mathcal{W}+$	-	-
StyleRig (Tewari et al., 2020b)	$\mathcal{W}+$	Self-supervised two-way cycle consistency	-
I2S++ (Abdal et al., 2020)	$(\mathcal{W}+, \mathcal{B})$	-	✓
BDInvert (Kang et al., 2021)	$(\mathcal{W}+, \mathcal{S})$	P-norm$^+$, Semantic consistency reg.	✓

Proposed Work. In this paper, we develop SPHIn*X* (StyleGAN with Projection Heads for Inverting *X*), an inversion approach for accurately embedding OOD images onto the latent space of StyleGAN-v2. We make a critical finding that, by redesigning the projection head that maps between $\mathcal{Z}+$ and $\mathcal{W}+$, such that the style latent variables corresponding to different intermediate layers in the generator architecture are decoupled, one can significantly improve the inverse optimization process. In a nutshell, SPHIn*X* improves OOD image inversion by: (a) replacing the existing mapping function f with a style projection head $\mathcal{P}_s$; (b) adopting a novel training strategy that enforces $\mathcal{P}_s$ to consistently produce meaningful solutions in the $\mathcal{W}+$ latent space for any realization from $P(\mathcal{Z}+)$; and (c) optimizing a content projection head $\mathcal{P}_c$ and the noise latent variables $\mathcal{B}$. We find that such a strategy results in a robust estimate of $P(\mathcal{W}+)$ for a given image.

Contributions. (a) A new approach, SPHIn*X* , for inverting OOD images onto the StyleGAN-v2 latent space; (b) Novel training strategy that induces a robust local neighborhood in $\mathcal{W}+$ for a given image; (c) Design of a style projection head that maps between $\mathcal{Z}+$ and $\mathcal{W}+$, to improve inversion with OOD data; (d) Extensive empirical studies on 'in-the wild' face and non-face image data to demonstrate the efficacy of SPHIn*X* in reconstruction, semantic editing and solving challenging inverse problems - denoising, compressed recovery and simultaneous inversion & attribute discovery (Appendix B); (e) Systematic study of the behavior of different existing latent space optimization strategies using a broad suite of image datasets; (f) Our codes are publicly accessible[2].

[2] https://github.com/Rakshith-2905/SPHInX

2. Background

GAN Inversion. This refers to the ill-posed problem of inferring a latent code or an embedding $\mathbf{z}$ for an image in the latent space of a pre-trained generative model $\mathcal{G}$. Such an inversion technique can be utilized for semantic manipulation or solving restoration tasks such as in-painting and compressed sensing (Bora et al., 2017). Projected gradient descent (PGD) (Abdal et al., 2019; Anirudh et al., 2019; Abdal et al., 2019; Shah & Hegde, 2018; Yeh et al., 2017; Raj et al., 2019) is a commonly adopted strategy, which optimizes for a latent vector that minimizes a discrepancy $\mathcal{L}(.,.)$ between the generated image $\mathcal{G}(\mathbf{z})$ and the given observation I. Mathematically,

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \ \mathcal{L}(\mathcal{G}(\mathbf{z}), I) + \mathcal{R}(.), \tag{1}$$

where $\mathcal{R}(.)$ is an additional regularizer. Common choices for $\mathcal{L}$ are: (i) pixel-wise mean squared error ($\mathcal{L}_{\text{MSE}}$) and (or) (ii) learned perceptual image patch similarity ($\mathcal{L}_{\text{LPIPS}}$) (Zhang et al., 2018) which is a perceptual similarity metric based on deep network activations (VGG-16 (Simonyan & Zisserman, 2015)). Mathematically,

$$\mathcal{L}_{\text{LPIPS}} = \sum_{\ell} \frac{1}{H_\ell W_\ell} \sum_{h,w,c} w_{\ell c} (\Psi^\ell_{I_{hwc}} - \Psi^\ell_{I'_{hwc}}))^2, \tag{2}$$

where (H_ℓ, W_ℓ) denotes the spatial size in layer ℓ and Ψ^ℓ denotes the ℓ^{th} latent layer of the adopted classifier. Further, $w_{\ell c}$ corresponds to the channel-level scaling vector, and I, I' are the images being compared.

StyleGAN Preliminaries. At its core, StyleGAN (Karras et al., 2019; 2017) relies on a mapping network f that transforms an input latent code $\mathbf{z} \in \mathbb{R}^{512}$ sampled from a Gaussian prior $P(\mathcal{Z})$ to a disentangled intermediate latent

code $\mathbf{w} \in \mathbb{R}^{512} \in \mathcal{W}$. The latent code $\mathbf{w}$ is then repeated $N_\ell = 18$ times and passed to each of the layers in $\mathcal{G}$ (Karras et al., 2019; Huang & Belongie, 2017). Differing from conventional generative models (Goodfellow et al., 2014; Radford et al., 2016), instead of directly passing $\mathbf{z}$ to the first layer, StyleGAN uses a constant input $\mathbf{s} \in \mathbb{R}^{4\times4\times512}$ (initially drawn at random from a Gaussian prior $P(\mathcal{S})$) which is progressively transformed in every layer with increasing resolution to synthesize the images. Additionally, StyleGAN employs a set of noise inputs sampled independently from a Gaussian prior $P(\mathcal{B})$, in every layer to improve the overall textural quality.

StyleGAN-based Inversion. Since pre-trained StyleGAN can be effectively leveraged as a prior for ill-posed image recovery and semantic editing, several StyleGAN-specific inversion studies have emerged recently (Abdal et al., 2019; 2020; Daras et al., 2021; Menon et al., 2020; Zhu et al., 2020b; Wulff & Torralba, 2020). While performing StyleGAN-based inversion, the choice of the latent space along with additional regularization techniques adopted become critical. Table 1 provides a comprehensive list of StyleGAN-based inversion strategies, along with their choice of latent space optimization and regularization.

Abdal et al. (*Image2StyleGAN*, or shortly I2S), first investigated the efficacy of inverting an image onto the intermediate $\mathcal{W}$ space of StyleGAN. They made a crucial observation that the reconstruction quality can be significantly improved by optimizing with an extended intermediate latent space $\mathcal{W}+ \subseteq \mathbb{R}^{N_\ell\times512}$, where every $\mathbf{w}^+ \in \mathcal{W}+$ was obtained by stacking N_ℓ realizations from $P(\mathcal{Z})$ using the mapping f. In addition, they demonstrated that $\mathcal{W}+$ offers a higher degree of freedom to guide the inversion compared to $\mathcal{W}$. As an extension, Abdal et al. identified that images can be reconstructed with improved granularity by optimizing the noise space $\mathcal{B}$ along with $\mathcal{W}+$ (*Image2StyleGAN++*, or shortly I2S++). Recently,Daras et al. proposed an inversion strategy that progressively included different layers of StyleGAN and optimized for latent codes in $\mathcal{Z}+$ that lie within an ℓ_1-ball around the manifold induced by the previous layer (*Intermediate Layer Optimization*, or shortly ILO). Here, every $\mathbf{z}^+ \in \mathcal{Z}+$ was obtained by stacking N_ℓ realizations from $P(\mathcal{Z})$. In this paper, we systematically study the behavior of different optimization strategies while inverting OOD images within the same domain (e.g., in-the-wild face images) as well as images from novel domains (e.g., medical images), and solving ill-posed restoration tasks.

3. Proposed Approach

3.1. Motivation

Broadly, our work is motivated by the need for (i) accurate inversion of images beyond the training set (Section 4, Appendix C); (ii) enabling improved ill-posed restoration of OOD images (Section 5); (iii) semantic editing in novel domains without requiring pre-specified encoders (Section 4, Appendix B); and (iv) re-purposing GAN priors for novel domains with limited data access, such as medical imaging (Section 5, Appendix C).

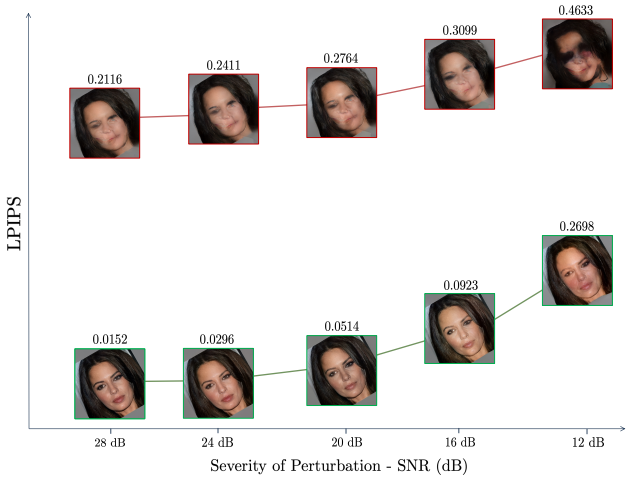

Figure 1: **Robustness of GAN inversion methods under latent space perturbations**. We show the perceptual quality of the reconstructed image (LPIPS defined in (2)) at different levels of noise perturbations (measured using signal-to-noise ratio). For in-the-wild face image with geometric transformation, the resulting solution ($(\mathcal{W}+, \mathcal{S}, \mathcal{B})$ in this illustration) is highly non-robust. In contrast, SPHIn*X* produces a solution that is perceptually more accurate as well as robust under perturbations.

While existing approaches can effectively invert images by optimizing in the extended $\mathcal{W}+$ latent space (along with semantic and noise latent variables $\mathcal{S}$ and $\mathcal{B}$), their performance with OOD images, e.g., geometrically altered or cartoonized faces, is found to be sub-optimal (Kang et al., 2021; Abdal et al., 2019). Given that the $\mathcal{W}+$ latent space contains rich semantic information about in-domain images (for e.g., faces), the latent codes need to be significantly altered in order to accurately reconstruct OOD images devoid of in-domain artifacts. Moreover, such an inversion is significantly challenging due to the inherently non-robust nature of $\mathcal{W}+$ space optimization. As illustrated using the LPIPS metric (from (2)) in Figure 1, the solution obtained by optimizing in the collection of latent spaces $(\mathcal{W}+, \mathcal{S}, \mathcal{B})$ of StyleGAN-v2 is highly non-robust even when a face image is rotated by only 30 degrees. Even minor perturbations to the solution (additive noise to achieve a target SNR) results in perceptually inferior reconstructions. This naturally motivates the need for novel optimization strategies that can infer solutions that are more locally robust for OOD images,

so that the inversion can stably converge. Due to lack of known priors on $\mathcal{W}+$, it is not straightforward to enforce such a local consistency during GAN inversion. In this paper, we address these shortcomings and obtain superior quality embeddings for OOD images through (i) optimization with carefully constructed projection heads (different from f); and (ii) a novel training paradigm that implicitly imposes a vicinal regularization in the $\mathcal{W}+$ space.

3.2. Optimization with Projection Heads

We propose to improve StyleGAN based inversion by optimizing a projection head that maps between $\mathcal{Z}+$ and $\mathcal{W}+$ instead of directly searching in either of the latent spaces via gradient descent. Intuitively, this crucial modification requires the inversion strategy to transform the prior distribution $P(\mathcal{Z}+)$ into an appropriate latent distribution $P(\mathcal{W}+)$, such that any realization from $P(\mathcal{W}+)$, when passed to the StyleGAN generator $\mathcal{G}$ will reconstruct the given image I. For instance, in the style latent space, the projection head $\mathcal{P}_s$ takes a realization from $P(\mathcal{Z}+)$, $\mathbf{z}^+ \in \mathcal{Z}+ \subseteq \mathbb{R}^{N_\ell \times 512}$ to produce a projected latent code $\mathbf{w}^+ \in \mathcal{W}+ \subseteq \mathbb{R}^{N_\ell \times 512}$.

A naïve way to implement this is to directly fine-tune the pre-trained mapping function f to perform OOD inversion, without directly manipulating the latent variables as done in all existing approaches. However, as shown in Figure 2, this results in poor quality embeddings and reconstructions $\hat{I}$ that do not contain any of the characteristics from the input image I. Furthermore, we also experimented with a randomly re-initialized f and found that this was also insufficient for inverting the image. This behavior can be attributed to the fact that, even with in-distribution images, the inversion can be improved only by individually controlling every latent vector in $\mathbf{w}^+$ (Richardson et al., 2021). Alternately, ILO (Daras et al., 2021) used a fixed mapping f, but adopted a novel optimization strategy that progressively included latent variables from different layers of StyleGAN and optimized for latent codes that lie within an ℓ_1-ball around the manifold induced by the previous layer. However, the inherent lack of local robustness for OOD images in the $\mathcal{W}+$ latent space makes such a progressive optimization also insufficient.

To circumvent this challenge, we design a *style projection head* $\mathcal{P}_s$ that decouples the latent spaces for different layers in $\mathcal{W}+$. In other words, $\mathcal{P}_s$ transforms each $\mathbf{z}^+ \in \mathcal{Z}+$ into $d-$dimensional representations using a *bottleneck* block of MLP layers. Subsequently, N_ℓ different decoder blocks (again a set of MLP layers) independently provide the corresponding mapping $\mathbf{w}^+ \in \mathcal{W}+$, using the bottleneck representation as input. Note that, while each $\mathbf{z}^+ \in \mathbb{R}^{512}$ and $\mathbf{w}^+ \in \mathbb{R}^{512}$, the choice of bottleneck dimension d is not very sensitive and we used $d = 16$ in all our experiments. Interestingly, using the proposed projection head and optimizing with different realizations of $\mathbf{z}^+$ from $P(\mathcal{Z}+)$ in every iteration of the optimization process, we obtain an accurate yet robust estimate of $P(\mathcal{W}+)$ for a given image. As illustrated in Figure 2, this results in a highly accurate reconstruction of in-the-wild face images using a StyleGAN-v2 pre-trained on FFHQ.

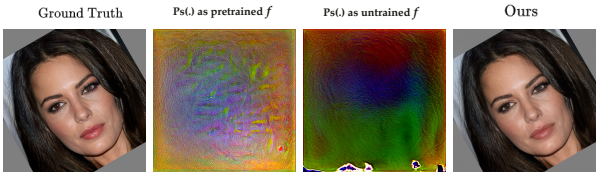

Figure 2: **Design of the projection head**. While re-purposing the pre-trained mapping function f from StyleGAN-v2 as the projection head fails completely. However, using the proposed projection head $\mathcal{P}_s$, which decouples the different latent spaces in $\mathcal{W}+$, leads to significantly higher quality reconstructions.

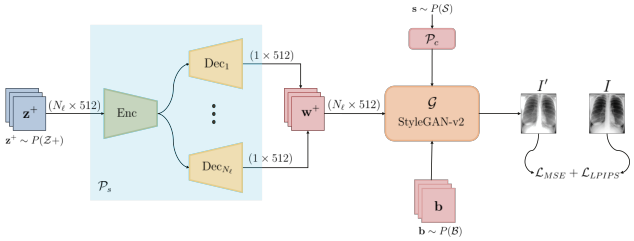

Figure 3: **Overview of SPHIn*X*** . For a given image I, SPHIn*X* trains a style projection head $\mathcal{P}_s$ (optionally along with a content projection head $\mathcal{P}_c$ and noise variables $\mathcal{B}$) through a novel training paradigm to effectively invert I onto the $\mathcal{W}+$ latent space of a pre-trained StyleGAN-v2.

As discussed earlier, the core idea of SPHIn*X* to improve the fidelity of OOD inversion is to perform optimization with the projection head $\mathcal{P}_s$. However, including content ($\mathcal{S}$) and noise ($\mathcal{B}$) latent parameters from StyleGAN-v2 into the collection of optimization variables (Wang et al., 2020; Kang et al., 2021; Daras et al., 2021) can further enhance the inversion fidelity. This is especially the case, when handling complex OOD shifts such as geometrically altered face and non-face images. Optimizing the content latent space $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^{4\times4\times512}$ (a constant tensor determined during GAN training) improves the inversion by better capturing the potentially unrelated semantic structure of OOD images. Based on our intuition on utilizing $\mathcal{P}_s$, we employ a content projection head $\mathcal{P}_c$ to directly parameterize the content input (randomly initialized using a Gaussian prior in lieu of the pre-trained content tensor). Further, optimizing the Gaussian noise inputs ($\mathcal{B}$) corresponding to each layer of the synthesis network in StyleGAN-v2 enhances the perceptual quality of the reconstructed images (Abdal

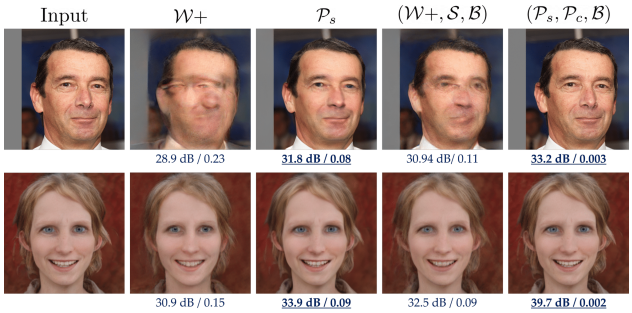

Figure 4: **Inversion of OOD images using different optimization strategies.** While SPHIn*X* ($\mathcal{P}_s$) improves the fidelity of reconstruction in both examples ((top) translated face; (bottom) cartoon), with the inclusion of additional optimization variables ($\mathcal{P}_c, \mathcal{B}$) SPHIn*X* significantly enhances the reconstruction quality (PSNR/LPIPS metrics are shown for every case).

Method	Translation				Rotation			Scaling			
	0	50	100	150	10	20	30	0.75	0.875	1.125	1.25
Image2StyleGAN	25.63	25.06	24.53	23.92	25.76	24.65	23.87	25.82	25.25	26.17	26.27
P-norm+	21.79	20.94	19.78	18.54	20.70	18.91	17.93	21.53	19.41	22.07	21.85
StyleGAN2 Inv.	18.73	18.29	17.31	16.71	17.95	17.22	16.02	18.65	18.43	19.12	19.43
PSP	20.54	19.03	17.59	16.50	19.14	17.78	16.99	19.02	17.78	20.63	20.15
BDInvert	26.47	26.30	26.37	26.43	26.48	26.49	26.33	26.44	26.28	26.98	27.26
SPHInX	**29.68**	**29.31**	**28.96**	**28.81**	**29.12**	**28.72**	**28.59**	**28.62**	**29.07**	**29.22**	**28.71**

Figure 5: **Quantitative comparison of in-the-wild face image reconstruction performance**. For this evaluation, we used the shifts: (i) translation (ii) rotation and (iii) scaling with varying levels of severity on in-the-wild faces. We utilize the PSNR metric for evaluating the reconstruction quality.

et al., 2020). Figure 3 illustrates a functional block diagram of our proposed approach. A detailed algorithm listing for SPHIn*X* is provided in the Appendix A.

4. Inverting In-the-Wild Face Images using StyleGAN-v2

The ability to accurately embed in-domain images (e.g. faces) onto the latent space of StyleGAN-v2 has enabled us solve a variety of challenging downstream tasks including image restoration, semantic editing and style transfer (Xia et al., 2021). However, when we consider misaligned images collected from the web or images characterized by unknown distribution shifts (or shortly "in-the-wild"), it is significantly challenging (Kang et al., 2021; Richardson et al., 2021) to obtain useful latent codes from a generator with semantic knowledge of only in-domain images. In this section, we demonstrate that SPHIn*X* can accurately recover such in-the-wild images with high-fidelity. Furthermore, we also find that our approach offers the ability to effectively interpolate between OOD face images and more importantly, manipulate specific attributes of interest (e.g., non-smiling $\rightarrow$ smiling), thus validating its utility in semantic editing and counterfactual reasoning (Axel Sauer, 2021).

4.1. In-the-Wild Face Image Reconstruction

In order to understand the impact of the optimization variables chosen for inversion, we consider OOD examples (translated face and cartoon images) and perform inversion (i) using SPHIn*X* ($\mathcal{P}_s$) and its variant ($\mathcal{P}_s, \mathcal{P}_c$ and $\mathcal{B}$) and (ii) using standard GAN inversion ($\mathcal{W}+$) and ($\mathcal{W}+, \mathcal{S}, \mathcal{B}$). Figure 4 provides the comparison between the reconstructions using these different choices. We can observe that through the effective re-parameterization of $\mathcal{W}+$, SPHIn*X* produces superior reconstructions in both cases over the baselines. However, the inclusion of $\mathcal{P}_c$ and $\mathcal{B}$ enhances the reconstruction quality. Based on this observation, all image reconstruction experiments reported in the paper involve the optimization of $\mathcal{P}_s, \mathcal{P}_c$ and $\mathcal{B}$ until otherwise specified.

Datasets. We evaluate the efficacy of SPHIn*X* for in-the-wild image inversion using face images collected from the web. We then applied different geometric transformations of varying severity levels to construct our evaluation set. Following Kang et al., we adopted the following domain shifts: (i) random rotations in the counterclockwise direction by 10, 20 and 30 degrees, (ii) scaling by factors of $7/8$, $3/4$, $9/8$, and $5/4$ and (iii) random translation by 0, 50, 100 and 150 pixels respectively. For all our experiments, we resized each image to a resolution of 1024×1024, and rescaled them to the range $[-1, 1]$.

Experiment Setup. The bottleneck block in $\mathcal{P}_s$ of SPHIn*X* is constructed using fully connected (FC) layers of sizes [512, 256, 16] and each of the decoders is another block of FCs of sizes [32, 64, 512]. The architecture for $\mathcal{P}_c$ is a fully convolutional network comprised of three Conv2D layers with $32, 128$ and 512 filters respectively. For each image, we trained all the compared methods for $10,000$ iterations using the ADAM optimizer with β_1 = 0.9 and β_2 = 0.999. We employ a trapezoid-based learning rate schedule with a maximum learning rate of 0.001. We used $N_\ell = 18$ corresponding to an image resolution of 1024×1024. For SPHIn*X* and the $\mathcal{Z}+$ baselines, we use a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^{N_\ell \times 512}$ as the prior for sampling $\mathbf{z}^+$. All approaches are trained using a combination of the mean-squared error (MSE) and LPIPS losses. For evaluating the quality of generated images in all experiments, we utilized two widely-adopted metrics: (i) *Peak Signal to Noise Ratio* (PSNR); (ii) *Learned Perceptual Image Patch Similarity* (LPIPS) (Zhang et al., 2018).

Figure 6: **Comparison of in-the-wild face image reconstruction performance**. In all cases, we optimize SPHIn*X* including $\mathcal{P}_c$ and $\mathcal{B}$. We find that even under complex geometric shifts, SPHIn*X* produces highly-accurate accurate reconstructions.

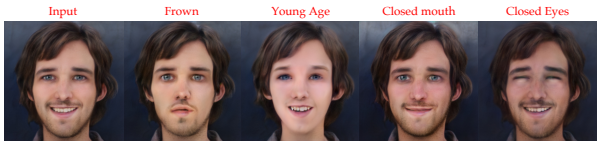

Figure 7: **Semantic editing of images from novel domains (cartoons)**. We observe that the solutions from SPHIn*X* can be manipulated using existing StyleGAN attribute directions when the OOD images are aligned with the FFHQ faces used for generator training.

Observations. Figure 6 provides a visual comparison of the images generated using SPHIn*X* along with the baselines across different geometric transformations. In contrast, strong baselines such as $(\mathcal{W}+,\mathcal{S},\mathcal{B})$ and ILO (equivalent to $(\mathcal{Z}+,\mathcal{S},\mathcal{B})$ when the ℓ_1-ball constraint is removed) fail to handle the unknown shifts. Figure 5 provides a quantitative comparison of SPHIn*X* on in-the-wild image reconstruction against several state-of-the-art baselines including Image2StyleGAN (Abdal et al., 2019), $\mathcal{P}-$norm(Zhu et al., 2020b), StyleGAN2 Inv (Karras et al., 2020) and the more recent PSP (Richardson et al., 2021) and BDInvert (Kang et al., 2021). It can be clearly seen that SPHIn*X* consistently produces higher PSNR in all cases.

4.2. In-the-Wild Face Image Editing

In addition to providing high-fidelity reconstructions, an important property of any GAN inversion approach is the ability to manipulate the image embeddings to synthesize plausible semantic changes. Given the utility of GAN inversion techniques in semantic editing tasks such as image diffusion and morphing (Xia et al., 2021), manipulating image embeddings along pre-specified StyleGAN directions reflective of unique attributes is a popular strategy to test the efficacy of inversion. If $\mathbf{w}_*^+$ denotes the solution obtained by inverting an image and $\mathbf{D}$ refers to a pre-specified direction vector, semantic editing can be performed as $\mathbf{w}^+ = \mathbf{w}_*^+ + \alpha\ \mathbf{D}$. Here α controls the intensity of attribute change. Interestingly, we find that SPHIn*X* supports such semantic editing operations for images even from novel domains that are geometrically aligned with FFHQ faces used for training the generator (e.g., cartoons). As demonstrated in Figure 7, existing attribute directions (e.g., age and closed mouth) in StyleGAN can be applied to solutions obtained from $\mathcal{P}_s$ for such images without requiring pre-trained encoders such as PSP (Richardson et al., 2021).

However, when handling complex OOD shifts such as geometrically altered face and non-face images that are misaligned with the FFHQ face image manifold, existing StyleGAN directions cannot be repurposed impeding semantic editing. In such scenarios, the ability to progressively synthesize meaningful realizations by interpolating between 'in-the wild' images that differ by one or more unknown attributes can be considered as an important evaluation benchmark. Given two OOD face images I_1 and I_2 that differ by an attribute (for e.g., smile), we first embed the images onto the StyleGAN-v2 latent space by optimizing $(\mathcal{P}_s,\mathcal{P}_c,\mathcal{B})$. Let θ_1 and θ_2 be the latent codes corresponding to the images I_1 and I_2. We now perform a linear interpolation

Figure 8: **Semantic interpolation between 'in the wild' images that differ by an unknown attribute (pose/smile/age).** SPHInX optimized along with $\mathcal{P}_c$ and $\mathcal{B}$ progressively traverses along the attribute and produces highly plausible realizations even for OOD images.

between the latent codes and obtain the corresponding realizations. Mathematically, $\theta_{int} = \alpha\ \theta_1 + (1-\alpha)\ \theta_2$ and $I_{int.} = \mathcal{G}(\theta_{int})$. Here, θ_{int} and $I_{int.}$ are the interpolated latent code and image respectively, and $\alpha \in [0, 1]$. Figure 8 illustrates the ability of SPHInX to progressively traverse along the semantic attribute, even under complex distribution shifts (30 degree rotation, 0.75 scaling and 150 pixel translation).

5. Ill-Posed Restoration of Non-face Images using a FFHQ StyleGAN-v2

GAN inversion with StyleGAN has been known to be effective at solving ill-posed image restoration tasks (Daras et al., 2021). While this can be a valuable tool in many applications, lack of access to pre-trained, high-quality generators or semantically rich, large-scale datasets is a critical limitation. In such scenarios, it becomes important to investigate if general-purpose generators from the vision community can be re-purposed to solve inverse problems with images from novel domains (e.g., medical imaging). In particular, the highly expressive StyleGAN latent spaces provide a natural avenue for inverting non-face images (while the GAN was trained only using faces).

Datasets. We consider a broad suite of non-face image datasets to evaluate the efficacy of SPHInX for ill-posed image inversion: (i) *Animal Faces-HQ (AFHQ)* (Choi et al., 2020): This dataset contains 16,130 high-quality images of various breeds of cats, dog and other wildlife; (ii) *Diabetic Retinopathy Images (Retina)* (ret): This dataset consists of 88,702 high-resolution, left and right eye retina images taken under a variety of imaging conditions; (iii) *ISIC 2018 Skin Lesions* (Codella et al., 2019): This dataset contains a total of 10,015 dermoscopic lesion images drawn from the HAM10000 database (Tschandl et al., 2018); and (iv) *Mimic CXR* (Johnson et al., 2019): This is a large public database containing 377,100 chest radiographs (X-rays) corresponding to a variety of radiographic studies; For all our experiments, we resized each image to a resolution of 1024×1024 and rescaled them to the range $[-1, 1]$. Following the observations from section 4, we optimize SPHInX along with $\mathcal{P}_c$ and $\mathcal{B}$ for all restoration experiments.

Image Reconstruction Evaluation. Figure 9 illustrates the image reconstruction performance of SPHInX for non-face images, and shows comparisons against baseline approaches. For each dataset, we show the median, along with the 25^{th} and the 75^{th} percentiles, of the metrics obtained from 50 randomly chosen images (Refer to the appendix for results on additional datasets). We find that SPHInX consistently outperforms the baselines across the two datasets, *AFHQ* and *Retina* thereby demonstrating its efficacy in OOD inversion. Interestingly, the performance of state-of-the-art approaches such as I2S++ (referred as $(\mathcal{W}+, \mathcal{S})$) and ILO are very similar across all metrics, and consistently lower than SPHInX . Comparatively, the $(\mathcal{W}+, \mathcal{S}, \mathcal{B})$ baseline is

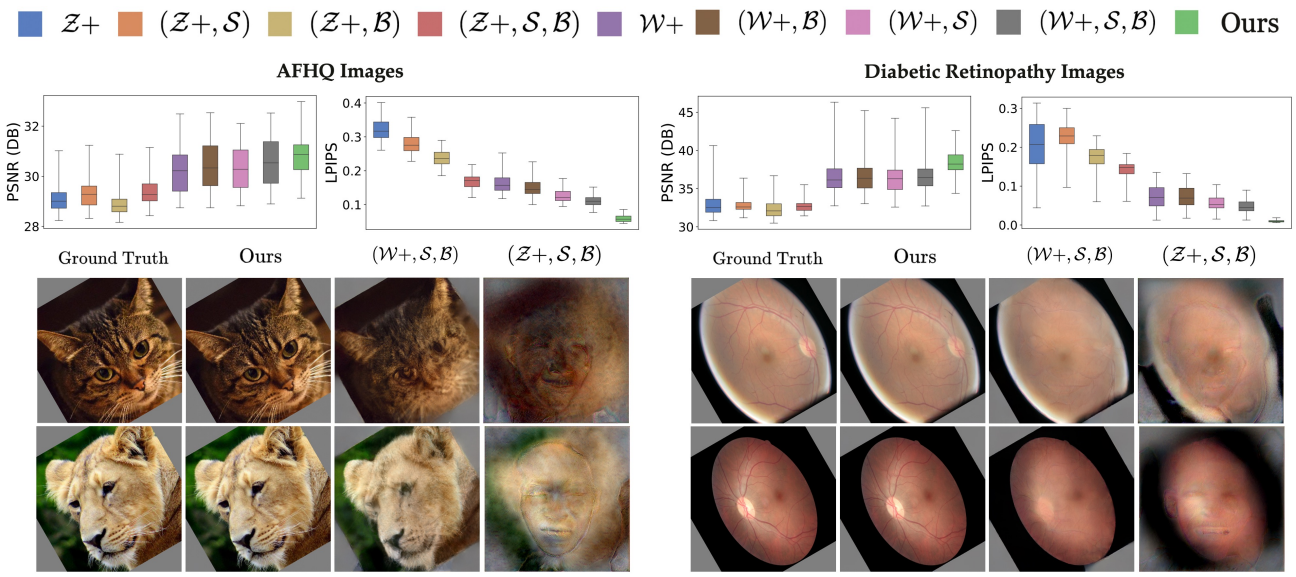

Figure 9: **Comparison of non-face image reconstruction performance**. Through the use of style and content projection heads, along with a novel training strategy, we find that SPHIn*X* consistently outperforms the baseline methods in both the metrics (LPIPS↓, PSNR↑) across the datasets. It must be noted that in all cases, we optimize SPHIn*X* along with $\mathcal{P}_c$ and $\mathcal{B}$.

the second best approach.

5.1. Denoising with a GAN Prior

Given an image corrupted by an unknown noise process, the goal is to restore the true image from the noisy observation. However, since noise, edges and textures are all high frequency components, it is challenging to effectively reconstruct an image without loss of details. Daras et al. demonstrated that StyleGAN can be used to remove noise while preserving details in the reconstructed image. Given the efficacy of SPHIn*X* in inverting non-face images with the FFHQ StyleGAN-v2, we now investigate its utility in denoising. For this experiment, we added Gaussian noise with known parameters and optimized our projection heads such the noise is suppressed.

Figure 10 shows the results of the denoising experiment, wherein the restored images from SPHIn*X* and the best-performing baseline $(\mathcal{W}+,\mathcal{S},\mathcal{B})$ are included. Furthermore, we also plot the metrics aggregated across 10 different examples and varying levels of corruption severity. More specifically, we show the mean and standard deviation for each of the metrics at each noise level. In particular, we measured the performance over increasing noise strengths (STD) in the range $[0.20, 1]$. It can be observed that SPHIn*X* consistently outperforms the baselines (higher PSNR and lower LPIPS), particular at lower noise level. However, at higher noise strengths, we find that there is a steep drop in the recovery performance, highlighting the limitation of the GAN prior in distinguishing true signal and noise in the images from an unseen domain.

5.2. Compressed Recovery of Medical Images

Compressed sensing is an image acquisition method, where we attempt to reconstruct images using a very few random measurements. This process offers compression of data below the Nyquist rate, which makes it an effective solution in the field of medical imaging, and has been extensively used for ultrasound (US) compression and sparse recovery. For this task, following (Daras et al., 2021), we generated observations of random projections using partial circulant measurement matrices with random signs and a fixed percentage of measurements. It must be noted that, we use the same measurement process, as that of the true image, with the StyleGAN-v2 reconstruction before evaluating the loss function. For the optimization, we used only the MSE loss between the generated and the true observations.

Figure 11 summarizes the results from the compressed recovery experiments with two medical imaging datasets, namely CXR and ISIC skin lesion. In addition to showing the PSNR and LPIPS metrics at different number of measurements (varied between 1% and 5%), we include examples of the images recovered using SPHIn*X* and the $(\mathcal{W}+,\mathcal{S},\mathcal{B})$ methods. We make a surprising finding even at severe compression factors, SPHIn*X* with an out-of-domain generator produces high-fidelity reconstructions (Average PSNR $\sim$ 34 dB) and convincingly outperforms existing baselines. We find that only SPHIn*X* is able to consistently recover finer details in the lesion images, wherein features

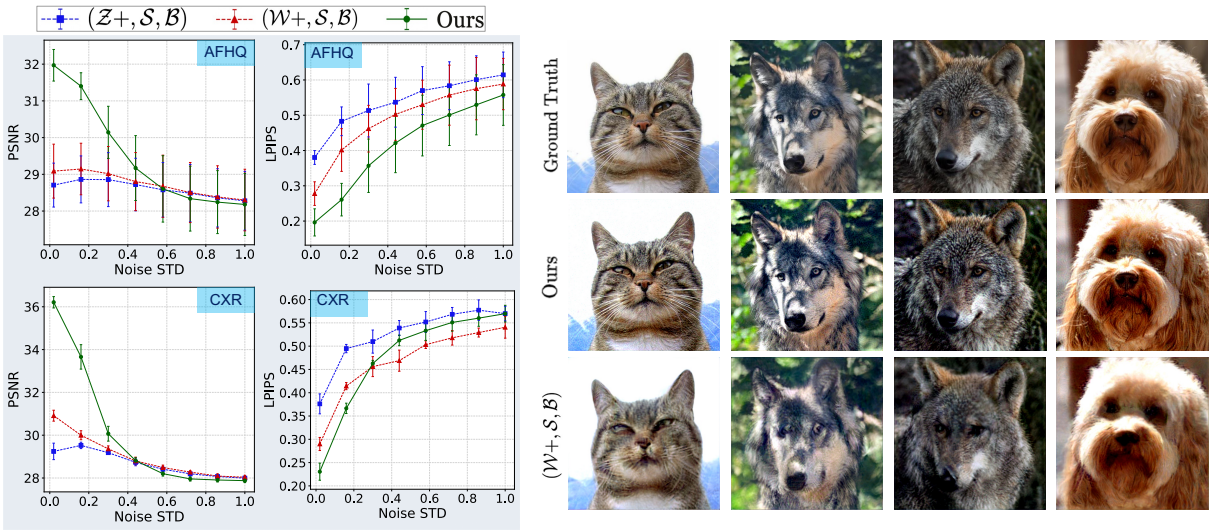

Figure 10: **Denoising non-face images using FFHQ StyleGAN-v2 prior**.(Left) shows that SPHIn*X* outperforms other baselines $(\mathcal{W}+,\mathcal{S},\mathcal{B})$ & $(\mathcal{Z}+,\mathcal{S},\mathcal{B})$ and produces high-quality images, particularly at lower noise levels; (Right) provides example reconstructions from SPHIn*X* & $(\mathcal{W}+,\mathcal{S},\mathcal{B})$ at noise std = 0.3.

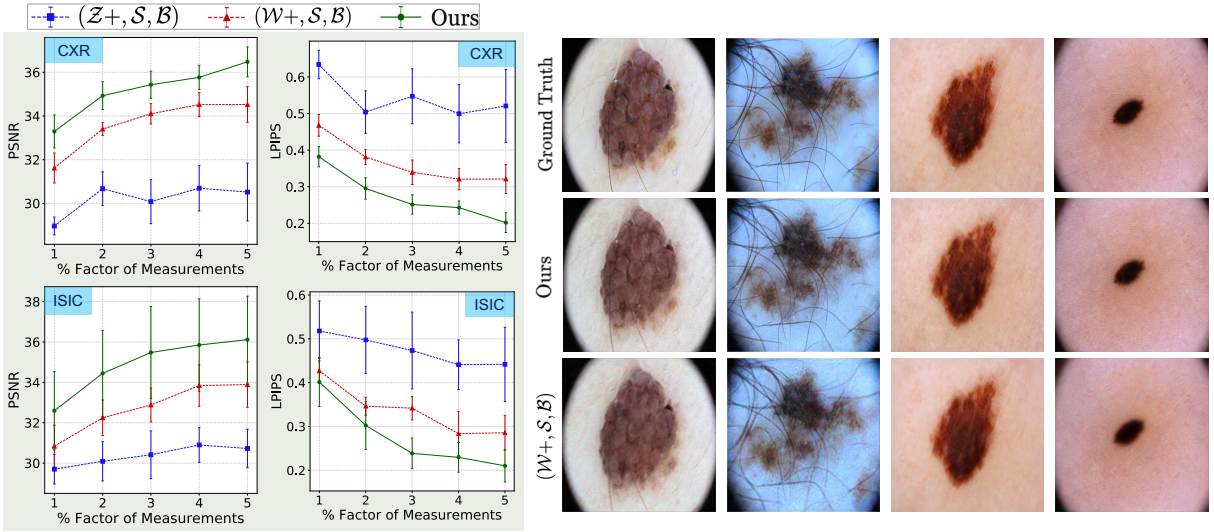

Figure 11: **Compressed recovery of medical images with a FFHQ StyleGAN-v2 prior**. (Left) shows SPHIn*X* consistently outperforms other baselines at different % factor of measurements. (Right) shows the performance of SPHIn*X* and $(\mathcal{W}+,\mathcal{S},\mathcal{B})$ when the percentage of measurements is as low as 1%.

such as the color and the blurriness along the edges are known to be crucial for making reliable diagnosis.

6. Conclusions

In this paper we presented SPHIn*X* , a new approach for solving ill-posed inverse problems with pre-trained StyleGAN-v2. Through the use of carefully designed projection heads for style and content latent spaces, and a novel training strategy, SPHIn*X* produces accurate and robust embeddings for even arbitrary OOD images. With extensive empirical studies with multiple datasets, we demonstrated significant performance improvements in embeddings high-resolution OOD images as well as ill-posed tasks such as denoising and compressed sensing. Compared to state-of-the-art approaches such as I2S++ (Abdal et al., 2020) and ILO (Daras et al., 2021), we find that SPHIn*X* stably converges to meaningful embeddings in the latent space. In summary, our study clearly evidences the utility of StyleGAN as a strong image prior even in domains, where collecting large datasets for training custom generative models is infeasible.

References

California healthcare foundation. diabetic retinopathy detection. URL https://www.kaggle.com/c/diabetic-retinopathy-detection. [Last accessed October 31, 2021]. 7

Abdal, R., Qin, Y., and Wonka, P. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019. 1, 2, 3, 6

Abdal, R., Qin, Y., and Wonka, P. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8296–8305, 2020. 1, 2, 3, 4, 9

Abdal, R., Zhu, P., Mitra, N. J., and Wonka, P. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 2

Anirudh, R., Thiagarajan, J. J., Kailkhura, B., and Bremer, T. Mimicgan: Robust projection onto image manifolds with corruption mimicking. *arXiv preprint arXiv:1912.07748*, 2019. 1, 2

Axel Sauer, A. G. Counterfactual generative networks. In *International Conference on Learning Representations (ICLR)*, 2021. 5

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *International Conference on Machine Learning*, pp. 537–546. PMLR, 2017. 1, 2

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm. 1

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 7

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 7

Daras, G., Dean, J., Jalal, A., and Dimakis, A. G. Intermediate layer optimization for inverse problems using deep generative models. *arXiv preprint arXiv:2102.07364*, 2021. 1, 2, 3, 4, 7, 8, 9

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 3

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9841–9850. Curran Associates, Inc., 2020. 1

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 1

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017. 3

Jahanian, A., Chai, L., and Isola, P. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. 1

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 7

Kang, K., Kim, S., and Cho, S. Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13941–13949, 2021. 1, 2, 3, 4, 5, 6

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019. 1, 2, 3

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020. 1, 6

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021. 1

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. Pulse: Self-supervised photo upsampling via latent space

exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020. 2, 3

Plumerault, A., Borgne, H. L., and Hudelot, C. Controlling generative models with continuous factors of variations. *arXiv preprint arXiv:2001.10238*, 2020. 1

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations*, 2016. 3

Raj, A., Li, Y., and Bresler, Y. Gan-based projector for faster recovery with convergence guarantees in linear inverse problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5602–5611, 2019. 2

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., and Cohen-Or, D. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2287–2296, 2021. 4, 5, 6

Shah, V. and Hegde, C. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4609–4613. IEEE, 2018. 2

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015. 2

Song, G., Luo, L., Liu, J., Ma, W.-C., Lai, C., Zheng, C., and Cham, T.-J. Agilegan: Stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, jul 2021. 1

Tewari, A., Elgharib, M., Bernard, F., Seidel, H.-P., Pérez, P., Zollhöfer, M., and Theobalt, C. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020a. 2

Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.-P., Pérez, P., Zollhofer, M., and Theobalt, C. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6142–6151, 2020b. 2

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018; sci data (5): 180161, 2018. 7

Voynov, A. and Babenko, A. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pp. 9786–9796. PMLR, 2020. 1

Wang, T., Zhang, Y., Fan, Y., Wang, J., and Chen, Q. High-fidelity gan inversion for image attribute editing. *arXiv preprint arXiv:2109.06590*, 2021. 1

Wang, Y., Gonzalez-Garcia, A., Berga, D., Herranz, L., Khan, F. S., and Weijer, J. v. d. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9332–9341, 2020. 4

Wulff, J. and Torralba, A. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020. 1, 2, 3

Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., and Yang, M.-H. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021. 5, 6

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5485–5493, 2017. 2

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 2, 5

Zhu, J., Shen, Y., Zhao, D., and Zhou, B. In-domain gan inversion for real image editing. In *European conference on computer vision*, pp. 592–608. Springer, 2020a. 2

Zhu, P., Abdal, R., Qin, Y., Femiani, J., and Wonka, P. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020b. 1, 2, 3, 6

Summary of the Appendices

Appendix	Title
A	Algorithm listing for SPHIn*X*
B	Additional Experiment: Simultaneous inversion and attribute discovery using SPHIn*X*
C	Non-face image reconstruction results for Mimic CXR and ISIC lesion datasets
D	Sensitivity of the Bottleneck Dimension d

A. Algorithm Listing for SPHIn*X*

Algorithm 1 provides the details of training SPHIn*X* [3].

Algorithm 1 SPHIn*X*

Input: Input image I,
No. of iterations N,
Learning rate η,
Pre-trained StyleGAN generator $\mathcal{G}$,
No. of generator layers N_ℓ,
Style Projection Head $\mathcal{P}_s(.;\theta)$,
Content Projection Head $\mathcal{P}_c(.;\phi)$,
Noises $\mathbf{b}$,
Penalties λ_1 for MSE and λ_2 for LPIPS.
Output: $\mathcal{P}_s(.;\theta^*)$, $\mathcal{P}_c(.;\phi^*)$, $\mathbf{b}^*$,
Generated Image I'.
Initialize: $P(\mathcal{Z}+) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \subseteq \mathbb{R}^{N_\ell \times 512}$,
$P(\mathcal{S}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \subseteq \mathbb{R}^{4\times4\times512}$.
for n in 1 to N **do**
  $\mathbf{z}^+ \sim P(\mathcal{Z}+)$; $\mathbf{s} \sim P(\mathcal{S})$
  Compute $\mathbf{w}^+ = \mathcal{P}_s(\mathbf{z}^+;\theta)$, $\mathbf{s}' = \mathcal{P}_c(\mathbf{s};\phi)$;
  Generate image $I' = \mathcal{G}(\mathbf{w}^+, \mathbf{s}', \mathbf{b})$;
  Compute loss $\mathcal{L}(I, I')$ using Eqn (1);
  $(\theta, \phi, \mathbf{b}) \leftarrow (\theta, \phi, \mathbf{b}) - \eta(\nabla_\theta\mathcal{L}, \nabla_\phi\mathcal{L}, \nabla_\mathbf{b}\mathcal{L})$;
end for
$\mathbf{w}^+ = \mathcal{P}_s(\mathbf{z}^+;\theta^*)$;
$\mathbf{s}' = \mathcal{P}_c(\mathbf{s};\phi^*)$;
return: $\mathcal{P}_s(.;\theta^*)$, $\mathcal{P}_c(.;\phi^*)$, $\mathbf{b}^*$, $I' = \mathcal{G}(\mathbf{w}^+, \mathbf{s}', \mathbf{b}^*)$

B. Additional Experiment: Simultaneous Inversion and Attribute Discovery for Novel Domain Images

A desired property of any GAN inversion algorithm is that the latent codes can be semantically manipulated for downstream applications such as style transfer and attribute discovery (Voynov & Babenko, 2020; Härkönen et al., 2020; Plumerault et al., 2020; Jahanian et al., 2019; Wang et al., 2021). However, when inverting images from novel domains (e.g., medical images onto the GAN latent space), the mismatch between the latent spaces and the OOD image makes it significantly hard to meaningfully manipulate the latent codes. Hence, we introduce a new inverse optimization problem to evaluate GAN inversion techniques in OOD settings. Given an image I along with its K variants that differ by a single attribute (for e.g., rotation), our goal is to simultaneously invert all $K+1$ images using a starting point $\mathbf{w}_*^+$ in the latent space for embedding I and local direction vectors in each of the layers, $\mathbf{D} \in \mathbb{R}^{N_\ell \times 512}$, along which the remaining K variants can be accurately

[3] SPHIn*X* uses publicly available StyleGAN-v2 pre-trained on FFHQ from https://github.com/rosinality/stylegan2-pytorch for all experiments.

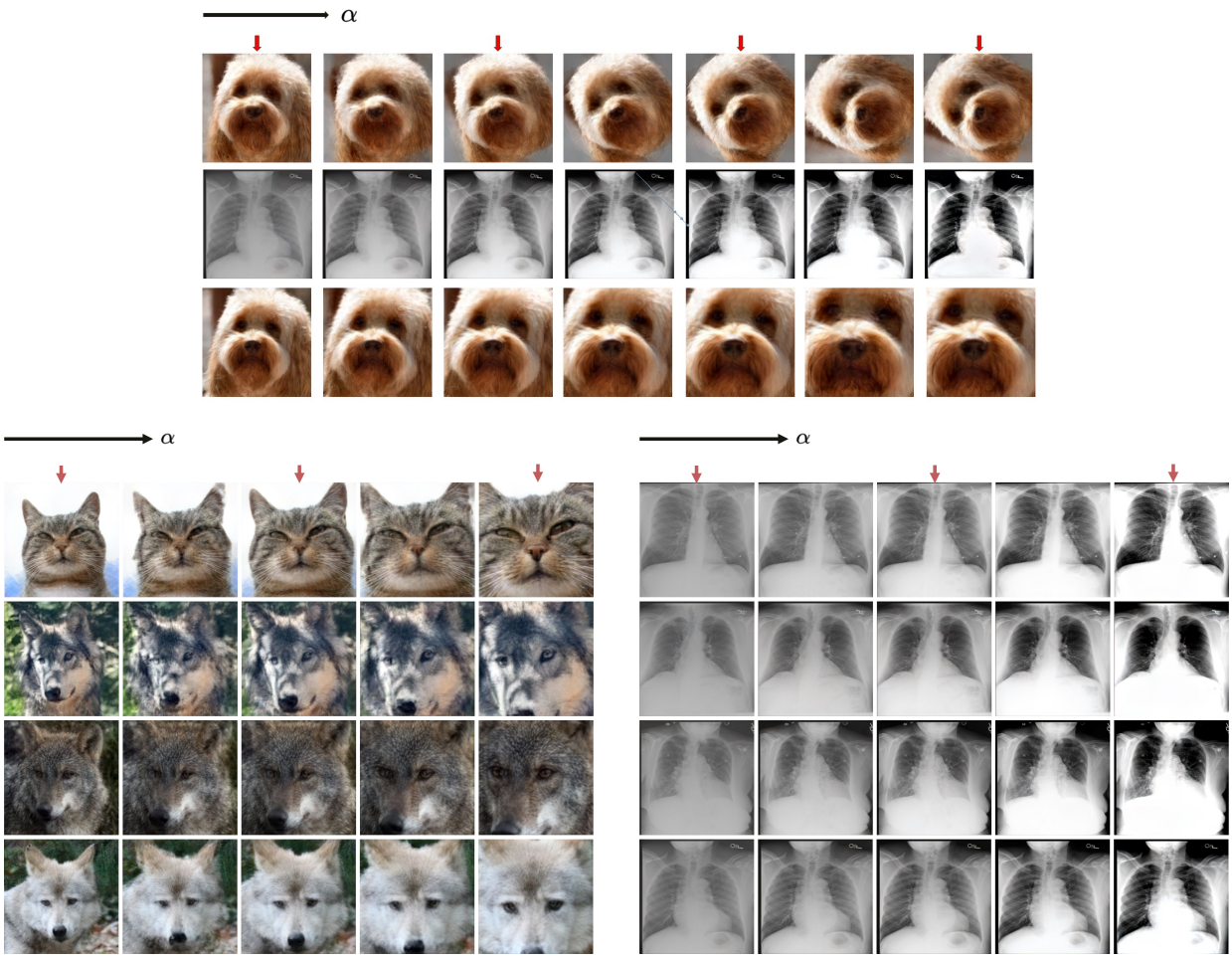

Figure 12: **Simultaneous image inversion and attribute discovery**. SPHIn*X* can learn meaningful attribute directions - rotation, brightness and zoom - by simultaneously inverting an image along with its realizations that differ by the attribute. By varying the scale of traversal α along the inferred direction, we observe that SPHIn*X* effectively produces realizations reflective of the learned attribute. In each case, the input images are marked with a red arrow.

embedded. Formally,

$$\min_{\mathbf{w}_*^+,\mathbf{D},\{\alpha\}} \mathcal{L}(\mathcal{G}(\mathbf{w}_*^+), I) + \sum_{k=1}^{K} \mathcal{L}\left(\mathcal{G}\left(\mathbf{w}_*^+ + \alpha_k \frac{\mathbf{D}}{\|\mathbf{D}\|_2}\right), I_k\right),$$

where $\{\alpha_k\}$ refers to the set of scaling parameters for each of the K images. Upon training, we expect the generator to synthesize manipulations pertinent to the learned attribute by traversing along $\mathbf{D}$ from $\mathbf{w}_*^+$.

Setup. In this study, we considered three different image transformations: (i) rotation, (ii) brightness and (ii) zoom and synthesized $K = 3$ different variants for each image by manipulating the chosen attribute. We adopt the following loss formulation in order to solve this optimization problem,

$$\begin{aligned} \mathcal{L}_{embed} \;=\; & \mathcal{L}(\mathcal{G}(\mathbf{w}_*^+), I) \;+ \\ & \sum_{k=1}^{K} \mathcal{L}\left(\mathcal{G}\left(\mathbf{w}_*^+ + \alpha_k \frac{\mathbf{D}}{\|\mathbf{D}\|_2}\right), I_k\right), \end{aligned}$$

where $\mathcal{L}$ denotes the weighted sum of the MSE and LPIPS losses. Here $\{\alpha_k\}_{k=1}^K$ denote the learnable scaling factors along the direction $\mathbf{D}$. It must be noted that $\mathbf{w}_*^+$, $\mathbf{D}$ and $\{\alpha_k\}_{k=1}^K$ are inferred during this optimization. In addition, we enforce $\{\alpha_k\}_{k=1}^K$ to be increasingly monotonic in an effort to map α_k to the intensity of the attribute change. Here, we consider

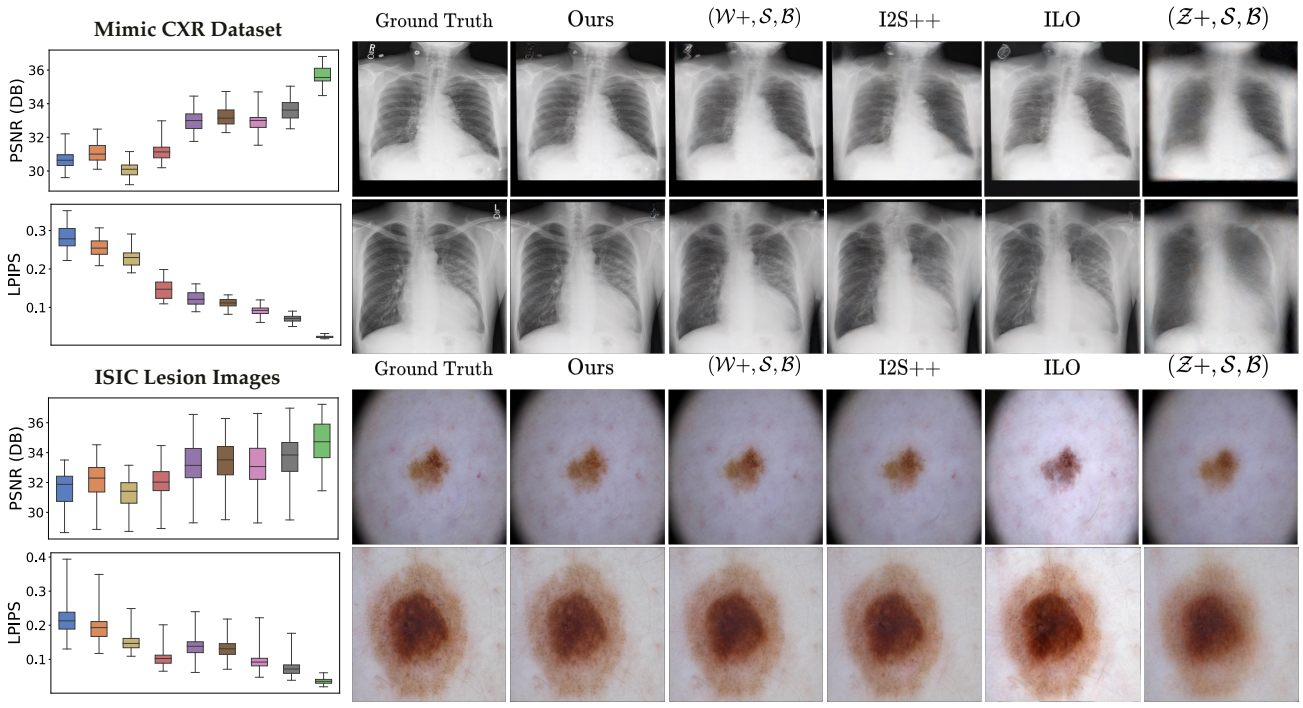

Figure 13: **Recovering non-face images using face GAN**. Using SPHIn*X* , we are able to effectively invert images from novel domains.

$\{I_k\}_{k=1}^{K}$ be arranged in an order of increasing attribute intensity change. We impose the constraint using a margin based loss function as follows,

$$\begin{aligned}\mathcal{L}_{margin} &= \max(0, \tau - \alpha_k) + \\ &\quad \max(0, \tau - \alpha_{k+1} + \alpha_k),\end{aligned}$$

where τ is the margin. The total optimization objective now becomes,

$$\mathcal{L}_{total} = \mathcal{L}_{embed} + \mathcal{L}_{margin}.$$

Training Details. We choose brightness, zoom and rotations as the attributes and generate $K = 3$ respective realizations of the given image I. We utilize the in-built, PyTorch transforms to generate the realizations. For the brightness attribute, we use brightness factors $\{0.5, 1.0, 1.5\}$ while for zoom, we use zoom in factors $\{1.0, 1.25, 1.75\}$. We generate K realizations corresponding to $\{0, 22.5, 45\}$ degrees for the rotation attribute. During inference, we vary the scale of traversal α along the learned attribute direction to generate the images. For all experiments, we use a margin $\tau = 2.0$.

Findings. Figure 12 illustrates the results from SPHIn*X* corresponding to all three attributes. The images are generated by traversing the learned direction vector $\mathbf{D}$ by varying α. Our results show that, SPHIn*X* can accurately recover directions corresponding to specific attribute changes in the StyleGAN latent space trained on FFHQ.

C. Non-face image reconstruction results for Mimic CXR and ISIC lesion datasets

In addition to the two datasets showed in Figure 9 for non-face image reconstruction using a StyleGAN-v2 trained on FFHQ, in Figure 13 we include the results for two additional medical imaging datasets. The observations on the improvements in terms of both PSNR and LPIPS holds in these cases as well.

D. Sensitivity of the Bottleneck Dimension d

The only requirement for the projection head $\mathcal{P}_s$ is to support obtaining independent W vectors for all 18 layers. Across different choices of $d \in \{16, 32, 64, 128\}$, we found the std. dev in reconstruction to be negligible (in the order of $0.1dB$ in PSNR, 0.0005 in LPIPS across different datasets).